AD-A080 885 WISCONSIN UNIV-MADISON DEPT OF STATISTICS F/G 12/1
THE LARGE SAMPLE BEHAVIOR OF TRANSFORMATIONS TO NORMALITY.(U)
JAN 79 F HERNANDEZ, R A JOHNSON N00014-78-C-0722
UNCLASSIFIED UWIS-DS-79-545 NL

1 OF 1
AD
A080885

END
DATE
FILMED
3 - 80
DDC

12 LEVEL II

15 N00014-78-C-0722

14 UWIS-DS-79-

9 Technical Report No. 545

12 21

11 January 1979

6 THE LARGE SAMPLE BEHAVIOR OF TRANSFORMATIONS TO NORMALITY

by

10 Fabián Hernández
University of Wisconsin

Richard A. Johnson
University of Wisconsin

The Large Sample Behavior of Transformations to Normality[1]

by

Fabian Hernandez[2]
University of Wisconsin

Richard A. Johnson
University of Wisconsin

Abstract

We investigate the large sample behavior of both the classical and Bayesian procedures for selecting a transformation to normality. The study of the large sample behavior clearly reveals the role played by the assumptions leading to the Box and Cox procedures. Based on our large sample results, we introduce an information number approach for transforming a known distribution to near normality. This latter procedure provides bench marks for the maximum possible amount of improvement through power transformations. We illustrate our procedure with three examples. Finally, we generalize our procedure to random vectors and linear models situations.

1. Introduction and Summary

The scale on which a variable is measured may not be the most appropriate for statistical analysis or describing variation. It may even happen that such a scale hides basic characteristics of the data. Nowadays, it is common practice to transform or re-express the data to uncover some of these basic properties. In the words of Tukey (1977) "We now regard re-expression as a tool, something to let us do a better job of grasping data."

Most statisticians are familiar with the marked improvement in normality that can be achieved by transforming a given data set. Our results help delineate the maximum possible amount of improvement in the sense that they correspond to unlimited sample size. A major step towards an objective way of determining a transformation was made by Box and Cox (1964). They

considered the parametric family of transformations

$$X^{(\lambda)} = \begin{cases} \frac{X^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log X, & \lambda = 0 \end{cases} \quad \text{for } X > 0 \qquad (1.1)$$

Box and Cox suggest both classical and Bayesian methods of finding a data-based power transformation that improves the validity of a full-rank normal linear model. We, however, focus our attention on the particular case of a random sample from a parent distribution with probability density function (p.d.f.) $g(\cdot)$. In this case, the Box and Cox method determines a transformations to normality based on a random sample of n positive random variables $X_1,\ldots,X_n$ with common p.d.f. $g(\cdot)$. Box and Cox make the critical assumption

Assumption 1.1 There exists a λ_t for which $X_i^{(\lambda_t)}$ is $N(\mu,\sigma)$ for some μ and σ.

Under Assumption 1.1, the p.d.f. of an untransformed observation is

$$f(x) = \frac{x^{\lambda_t - 1}}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}[x^{(\lambda_t)}-\mu]^2\right\} \qquad (1.2)$$

for $x < -1/\lambda_t$ if $\lambda_t < 0$, $x > -1/\lambda_t$ if $\lambda_t > 0$ and $-\infty < x < \infty$ if $\lambda_t = 0$. Thus, the log-likelihood function in terms of the original observations is

$$\ell(\underset{\sim}{\theta}_t|\underset{\sim}{x}_n) = -\frac{n}{2}\,\ell og(2\pi) - n\,\ell og(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[x_i^{(\lambda_t)}-\mu]^2 + (\lambda_t - 1)\sum_{i=1}^{n}\ell og(x_i) \qquad (1.3)$$

where $\underset{\sim}{\theta}'_t = (\mu,\sigma,\lambda_t)$ and $\underset{\sim}{x}'_n = (x_1,\ldots,x_n)$. Box and Cox suggest using $\hat{\lambda}_t$ for the transformation where $\hat{\theta}_n = (\hat{\mu},\hat{\sigma},\hat{\lambda}_t)'$ is the MLE that maximizes (1.3).

For their Bayesian analysis, Box and Cox assume that the conditional distribution of $X^{(\lambda)}$ given $\underset{\sim}{\theta} = (\mu,\sigma,\lambda)'$ is normal so

[1] This research was sponsored by the Office of Naval Research under Grant No. N00014-78-C-0722.

[2] Now at Direccion General de Estadística, México.

$$p(\underset{\sim}{x}_n|\underset{\sim}{\theta}) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{-\frac{\nu S^2(\lambda)+n[\hat{\mu}(\lambda)-\mu]^2}{2\sigma^2}\right\} \prod_{i=1}^{n} x_i^{\lambda-1} \tag{1.4}$$

Where $S^2(\lambda) = \sum_{i=1}^{n} [x_i^{(\lambda)}-\hat{\mu}(\lambda)]^2 / (n-1)$, $\hat{\mu}(\lambda) = \sum_{i=1}^{n} x_i^{(\lambda)}/n$ and $\nu=n-1$.

In order to find the marginal posterior density of λ , Box and Cox utilize the data-based joint prior proportional to

$$p(\lambda) \frac{d\mu d\sigma d\lambda}{\sigma(\prod_{i=1}^{n} x_i)^{\frac{\lambda-1}{n}}} \tag{1.5}$$

and the marginal posterior density becomes

$$p(\lambda|\underset{\sim}{x}_n) = \frac{K_n\Gamma(\frac{\nu}{2})}{2\sqrt{\pi}(\nu\pi)^{\frac{\nu}{2}}} \frac{(\prod_{i=1}^{n} x_i^{\lambda-1})^{\frac{\nu}{n}}}{[S^2(\lambda)]^{\frac{\nu}{2}}} \quad p(\lambda) \tag{1.6}$$

where K_n is a positive constant, which does not depend on λ, such that $\int p(\lambda|\underset{\sim}{x}_n)d\lambda = 1$.

In section 2, we investigate the asymptotic behavior of the MLE $\hat{\lambda}_t$ and the mode, $\tilde{\lambda}_n$, of the marginal posterior distribution (1.6) of λ. It is worth noting that the mode of the joint posterior distribution of θ does not occur at the MLE $(\hat{\mu}(\hat{\lambda}_t), \hat{\sigma}(\hat{\lambda}_t), \hat{\lambda}_t)$ when the prior density is (1.5). Our development shows the relevance of the underlying 'true' distribution, which may not be transformable to an exact normal distribution. In particular, the true distribution determines the limiting variance of the MLE $\hat{\lambda}_t$.

Employing the Kullback-Leibler information number as a measure of discrepancy between two distributions, we present in Section 3 a method of transforming a random variable (r.v.) with known distribution to normality. This provides a new view of the large sample value of the transformation parameter λ selected by the Box-Cox method. Moreover, our approach allows us to measure both numerically and graphically the improvement to normality introduced by the transformation. We also consider two examples that show how a power transformation can produce approximate normality.

*A short discussion on the transformation of multivariate data, high*lighting transformations of marginal distributions and the assumption of a diagonal covariance matrix, appears in Section 4. We conclude by indicating how our earlier results, concerning the limiting behavior of the estimator of the power parameter and the information number, extend to linear model applications.

2. Asymptotic Results

Recall that the log-likelihood function (1.3) was derived by pretending that there exists a value of $\underset{\sim}{\theta}$, $\underset{\sim}{\theta}_0' = (\mu_0,\sigma_0,\lambda_0)$ for which the distribution of $X^{(\lambda_0)}$ is normal with mean μ_0 and standard deviation σ_0. Except for the log-normal case, $X^{(\lambda_0)}$ cannot be normal for positive random variables. We now show the consequence of maximizing the wrong log-likelihood function (1.3).

Draper and Cox (1969) tried to derive properties of $\hat{\lambda}_0$ but Hinkley (1975) found errors in their derivations that invalidate some of their results. Moreover, Hinkley stated, under rather loose conditions, a theorem giving the asymptotic distribution of the MLE $\hat{\underset{\sim}{\theta}}_n$. The purpose of Theorem 2.2 below is to formalize his result by giving conditions under which the log-likelihood function (1.3) satisfies a uniformity condition so the MLE $\hat{\underset{\sim}{\theta}}_n$ is strongly consistent and has an asymptotic normal distribution. The proof of Theorem 2.2 requires some properties of both the transformation $X^{(\lambda)}$ and the log-likelihood (1.3).

Lemma (2.1) Define $\phi:(0,\infty)X(-\infty,\infty) \to (-\infty,\infty)$ as

$$\phi(x,\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ell og(x), & \lambda = 0 \end{cases}$$

Then

1) $\phi(x,\lambda) > 0$ if $x > 1$ and $\phi(x,\lambda) \leq 0$ if $0 < x \leq 1$
2) $\phi(\cdot,\cdot)$ is increasing in both variables
3) $\phi(\cdot\ \cdot)$ is convex in λ for $x \geq 1$ and concave in λ for $x \leq 1$.
4) $\frac{\partial^r}{\partial\lambda^r}\phi(x,\lambda)$ is continuous in x and λ, $r \geq 1$.

Furthermore, let Θ be defined by (2.1) below and let $g(\cdot)$ be such that $E_g(X^{2a})$, $E_g(X^{2b})$, $E_g[X^a\ \ell og(X)]^2$ and $E_g[X^b\ \ell og(X)]^2$ are finite. $E_g(\cdot)$ means that the expected value is being taken with respect to the p.d.f. $g(\cdot)$. Then, with $\ell(\underset{\sim}{\theta}|X)$ given by (1.3), the random variables

$$\left|\frac{\partial^2\ell(\underset{\sim}{\theta}|X)}{\partial\theta_i\partial\theta_j}\right| , \quad i,j = 1,2,3 \quad , \quad \underset{\sim}{\theta}' = (\theta_1,\theta_2,\theta_3) = (\mu,\sigma,\lambda)$$

are dominated in absolute value by g-integrable functions for all $\underset{\sim}{\theta} \in \Theta$.

Proof A proof of this lemma can be found in Hernández (1978).

Theorem 2.2 Suppose the parameter space Θ, the true p.d.f. $g(\cdot)$ and the log-liklihood function (1.3) satisfy the following conditions

i) The parameter space Θ is a compact set defined as

$$\Theta = \{\theta = (\mu,\sigma,\lambda)' \,|\, |\mu| \leq M,\ c \leq \sigma \leq d,\ a \leq \lambda \leq b \text{ with } -\infty < a < 0 < b, d, c, M < \infty\}. \qquad (2.1)$$

ii) The true p.d.f. $g(\cdot)$ is concentrated on $(0,\infty)$ and the moments $E_g(X^{2a})$ and $E_g(X^{2b})$ are finite.

iii) $E_g[\ell(\underset{\sim}{\theta}|X)]$ has a unique global maximum at $\underset{\sim}{\theta}_0$.

Then

1) $\lim_{n\to\infty} \max_{\underset{\sim}{\theta}\in\Theta}[\frac{1}{n}\ \ell(\underset{\sim}{\theta}|\underset{\sim}{X}_n)]\} = \max_{\underset{\sim}{\theta}\in\Theta} E_g[\ell(\underset{\sim}{\theta}|X)]\}$, with probability one.

2) The MLE $\hat{\underset{\sim}{\theta}}_n$ is a strongly consistent estimator of $\underset{\sim}{\theta}_0$.
That is, $\hat{\underset{\sim}{\theta}}_n \to \underset{\sim}{\theta}_0$ as $n \to \infty$, with probability one.

Furthermore, if

iv) $\underset{\sim}{\theta}_0$ is an interior point of Θ

v) $E_g[X^a\ell og(X)]^2$ and $E_g[X^b\ell og(X)]^2$ are finite.

vi) $E_g[\nabla\ell(\underset{\sim}{\theta}_0|X)] = \underset{\sim}{0}$, where the column vector

$$\nabla\ell(\underset{\sim}{\theta}_0|X) = \left(\frac{\partial\ell(\underset{\sim}{\theta}|X)}{\partial\theta_i}\bigg|_{\underset{\sim}{\theta}=\underset{\sim}{\theta}_0}\right)$$ is the gradient of the log-likelihood function for $\underset{\sim}{\theta}' = (\theta_1,\theta_2,\theta_3) = (\mu,\sigma,\lambda)$.

vii) $E_g[\nabla^2\ell(\underset{\sim}{\theta}_0|X)]$ is non-singular, where $\nabla^2\ell(\underset{\sim}{\theta}_0|X)$

$$= \left(\frac{\partial^2\ell(\underset{\sim}{\theta}|X)}{\partial\theta_i\partial\theta_j}\bigg|_{\underset{\sim}{\theta}=\underset{\sim}{\theta}_0}\right)$$ is the Hessian of the log-likelihood function.

Then

3) $\sqrt{n}(\hat{\underset{\sim}{\theta}}_n-\underset{\sim}{\theta}_0) \xrightarrow{d} N_3(\underset{\sim}{0},VWV')$ where $V = \{E_g[\nabla^2\ell(\underset{\sim}{\theta}_0|X)]\}^{-1}$

and $W = E_g\{\nabla\ell(\underset{\sim}{\theta}_0|X)[\nabla\ell(\underset{\sim}{\theta}_0|X)]'\}$.

The proof is given in the Appendix.

Remark The importance of the true p.d.f. is reflected in the expression of the asymptotic covariance matrix of $\hat{\underset{\sim}{\theta}}_n$. On a minor detail, Hinkley (p.105) incorrectly states that $V=W^{-1}$ if $g(\cdot)$ (he uses f) is the normal p.d.f. . What is true is that $V = W^{-1}$ only if $g(\cdot)$ is the p.d.f. of a log-normal distribution. □

At first sight the assumption vi), that asserts $E_g[\nabla\ell(\underset{\sim}{\theta}_0|X)] = \underset{\sim}{0}$, may appear to be restrictive. In Hernández (1978) it is shown to hold for the log-normal, Gamma, Weibull, Inverse Gaussian and Pareto distributions.

We now turn our attention to the asymptotic behavior of the marginal posterior mode $\tilde{\lambda}_n$, when we use the joint prior (1.5) restricted to the parameter set (2.1). Although Berk (1966) has already studied the convergence of posterior distributions to a degenerate distribution, we will derive our result directly, to avoid modifying his results to include data dependent prior distributions.

Theorem 2.3 Let $p(\lambda)$, the prior p.d.f. of λ, be positive and continuous on the interval $[a,b]$ where $-\infty<a<0<b<\infty$. Suppose $E_g(X^{2a})$ and $E_g(X^{2b})$ are finite and that we use the joint prior (1.5) restricted on the set (2.1). Then, if $E_g[\ell(\underset{\sim}{\theta}|X)]$ has a unique global maximum at $\underset{\sim}{\theta}_0 = (\mu_0,\sigma_0,\lambda_0)'$, with $\underset{\sim}{\theta}_0$ in the set (2.1), then $\tilde{\lambda}_n \to \lambda_0$ with probability one.

The proof is given in the Appendix.

So far, we have seen that the M L E $\hat{\lambda}_n$ and the posterior mode $\tilde{\lambda}_n$ converge, under suitable conditions, to λ_0 the Normal Theory Value of λ where $\underset{\sim}{\theta}_0' = (\mu_0,\sigma_0,\lambda_0)$ is characterized by the inequality $E_g[\ell(\underset{\sim}{\theta}|X)] \leq E_g[\ell(\underset{\sim}{\theta}_0|X)]$ for all $\underset{\sim}{\theta}\in\Theta$. In the next section, we provide a new meaning to λ_0, the value of λ selected, asymptotically, by the Box-Cox technique.

3. A New View of the Transformation Parameter

Let X be a positive r.v. with known p.d.f. $g(\cdot)$. Let $f_\lambda(\cdot)$ be the p.d.f. of the transformed variable $U = X^{(\lambda)}$ and $\phi_{\mu\sigma}(\cdot)$ be the p.d.f. of a normal distribution with mean μ and standard deviation σ.

As a measure of discrepancy between two probability distributions, we use the Kullback-Leibler information number (see Kullback (1968)).

Definition Let h_1 and h_2 be two absolutely continuous candidate p.d.f.'s for a r.v. Z. The quantity

$$I[h_1,h_2] = \int h_1(t)\ell og\left\{\frac{h_1(t)}{h_2(t)}\right\} dt \tag{3.1}$$

is called the Kullback-Leibler information number. □

Now, suppose we are interested in transforming X so that the distribution of $X^{(\lambda)}$ is approximately normal. We propose to select the transformation λ so that $I[f_\lambda;\phi_{\mu\sigma}]$ is minimized for suitable choices of μ,σ and λ.

Proposed Procedure Select λ so that the information number between $f_\lambda(\cdot)$ and $\phi_{\mu\sigma}(\cdot)$ is a minimum. That is

$$\underset{\mu,\sigma,\lambda}{\text{minimize}} \int f_\lambda(u)\ \ell og\left\{\frac{f_\lambda(u)}{\phi_{\mu\sigma}(u)}\right\} du \tag{3.2}$$

Heuristically, the best choice of λ will produce an f_λ that is hardest to discriminate from, or is 'closest' to, a normal p.d.f. Analytically, we find it convenient to first find the values of μ and σ that best approximate $\phi_{\mu\sigma}$ by f_λ and then to search for the value of λ that minimizes the remaining 'distance'. Our next result provides the values of μ and σ so that $\phi_{\mu\sigma}$ is 'closest' to f_λ when λ is fixed.

Lemma 3.1 Let λ be fixed and $X^{(\lambda)}$ defined as in (1.1). Assume that the expected values $E_g(X^{2\lambda})$, $E_g[(\ell og\ X)^2]$ are finite. Then, the values of μ and σ that minimize $I[f_\lambda;\phi_{\mu\sigma}]$ are

$$\mu_*(\lambda) = E_g(X^{(\lambda)})$$

and

$$\sigma_*^2(\lambda) = E_g[X^{(\lambda)} - E_g(X^{(\lambda)})]^2 \tag{3.3}$$

Proof Since by assumption $E_g(X^{2\lambda})$ and $E_g[\ell og(X)]^2$ are finite it follows that $E_{f_\lambda}\{\ell og[\phi_{\mu\sigma}(U)]\} < \infty$. Also, because $f_\lambda(x^{(\lambda)}) = x^{1-\lambda}\ g(x)$

$$\begin{aligned} I[f_\lambda;\ \phi_{\mu\sigma}] &= E_{f_\lambda}\ \{\ell og[f_\lambda(U)]\} - E_{f_\lambda}\ \{\ell og[\phi_{\mu\sigma}(U)]\} \\ &= E_g\{\ell og[g(X)]\} + (1-\lambda)E[\ell og(X)] - E_g\{\ell og[\phi_{\mu\sigma}(X^{(\lambda)})]\} \end{aligned} \tag{3.4}$$

Setting $G(\lambda) = \min\limits_{\mu,\sigma^2} I[f_\lambda;\ \phi_{\mu\sigma}]$ and $V_g(X^{(\lambda)}) = E_g[X^{(\lambda)} - E_g(X^{(\lambda)})]^2$ then

$$G(\lambda) = \text{const} + (1-\lambda)E_g[\ell og(X)] + \frac{1}{2}\min_{\mu,\sigma^2}\left\{\ell og(\sigma^2) + \frac{V_g(X^{(\lambda)})}{\sigma^2} + \frac{[E_g(X^{(\lambda)})-\mu]^2}{\sigma^2}\right\} \tag{3.5}$$

$$\geq \text{const} + (1-\lambda)E_g[\ell og(X)] + \frac{1}{2}\{1 + \ell og[V_g(X^{(\lambda)})]\} \tag{3.6}$$

where $\text{const} = E_g\{\ell og[g(X)]\} + \frac{1}{2}\ \ell og(2\pi)$. The lower bound (3.6) is achieved by $\mu = E_g(X^{(\lambda)}) = \mu_*(\lambda)$ and $\sigma^2 = V_g(X^{(\lambda)}) = \sigma_*^2(\lambda)$ and these values match the population moments determined from $g(\cdot)$. □

For the choices $\mu_*(\lambda)$ and $\sigma_*^2(\lambda)$, the function $G(.)$ becomes

$$G(\lambda) = \frac{1}{2}[\ell og(2\pi)+1] + E_g\{\ell og[g(X)]\} + (1-\lambda)E_g[\ell og(X)] + \frac{1}{2}\ \ell og[V_g(X^{(\lambda)})]. \tag{3.7}$$

Thus, λ_* the optimal value of λ is found by minimizing $G(\cdot)$. It is clear from (3.7) that the selection of λ is scale invariant and so the selection of λ_* is independent of scale parameters.

Remark We want to make explicit the fact that the sequential minimization of $I[f_\lambda;\phi_{\mu\sigma}]$ does yield a global minimum. Suppose $\underset{\sim}{\theta}_1 = (\mu_1,\sigma_1,\lambda_1)'$ minimizes $I[f_\lambda;\phi_{\mu\sigma}]$. Then $I[f_{\lambda_1};\phi_{\mu_1\sigma_1}]$ $\leq I[f_{\lambda_*};\phi_{\mu_*(\lambda_*)\sigma_*(\lambda_*)}] = \min\limits_\lambda\ I[f_\lambda;\phi_{\mu_*(\lambda)\sigma_*(\lambda)}] \leq I[f_{\lambda_1};\phi_{\mu_*(\lambda_1)\sigma_*(\lambda_1)}]$ $\leq I[f_{\lambda_1};\phi_{\mu_1\sigma_1}]$ by Lemma (3.1). Therefore $I[f_{\lambda_1};\phi_{\mu_1\sigma_1}]$ $= I[f_{\lambda_*};\phi_{\mu_*(\lambda_*)\sigma_*(\lambda_*)}]$. □

We now consider the relation between λ_*, the value of λ that minimizes $I[f_\lambda;\phi_{\mu\sigma}]$, and λ_0, the limiting value of the Box-Cox MLE. We state our result as

Theorem 3.2 Let $X_1,\ldots,X_n$ be positive i.i.d.r.v.'s with p.d.f. $g(\cdot)$. Suppose we use the log-likelihood function (1.3) to estimate $\underset{\sim}{\theta}' = (\mu,\sigma,\lambda)$ and the assumptions i),ii) and iii) of Theorem (2.2) hold. Then, the MLE $\hat{\underset{\sim}{\theta}}_n$ satisfies $\hat{\underset{\sim}{\theta}}_n \to \underset{\sim}{\theta}_* = (\mu_*,\lambda_*,\sigma_*)' = (\mu_*(\lambda_*),\sigma_*(\lambda_*),\lambda_*\ \ )'$ with probability one and $\underset{\sim}{\theta}_*$ is the value of $\underset{\sim}{\theta}$ that minimizes $I[f_\lambda;\phi_{\mu\sigma}]$ gi·en by (3.2).

Moreover,

$$\underset{\sim}{\theta}_* = \underset{\sim}{\theta}_0$$

where $\underset{\sim}{\theta}_0$ is the normal theory value of $\underset{\sim}{\theta}$ in assumption iii) of Theorem (2.2).

Proof By (3.4)

$$I[f_\lambda;\phi_{\mu\sigma}] = E_g\{\ell og[g(X)]\} - E_g\{\ell og[\phi_{\mu\sigma}(X^{(\lambda)})X^{\lambda-1}]\}$$

hence

$$\min_{\underset{\sim}{\theta}\in\Theta} I[f_\lambda;\phi_{\mu\sigma}] = E_g\{\ell og[g(X)]\} - \max_{\underset{\sim}{\theta}\in\Theta} E_g\{\ell og[\phi_{\mu\sigma}(X^{(\lambda)})X^{\lambda-1}]\} \ .$$

Now, by Theorem (2.2), with probability one

$$\begin{aligned}\lim_{n\to\infty} \{\max_{\underset{\sim}{\theta}\in\Theta}[\tfrac{1}{n}\ell(\underset{\sim}{\theta}|\underset{\sim}{X}_n)]\} &= \max_{\underset{\sim}{\theta}\in\Theta} E_g[\ell(\underset{\sim}{\theta}|X)] \\ &= \max_{\underset{\sim}{\theta}\in\Theta} E_g\{\ell og[\phi_{\mu\sigma}(X^{(\lambda)})X^{\lambda-1}]\} \ .\end{aligned}$$

where the last equality follows from (1.3) evaluated for $n = 1$. Thus, the value of $\underset{\sim}{\theta}$ in Θ that maximizes $E_g[\ell(\underset{\sim}{\theta}|X)]$, or the normal value, is the same one that minimizes $I[f_\lambda;\phi_{\mu\sigma}]$. Therefore, $\hat{\underset{\sim}{\theta}}_n \to \underset{\sim}{\theta}_*$ with probability one and $\underset{\sim}{\theta}_0 = \underset{\sim}{\theta}_*$. □

Remark Theorem 3.2 says that the MLE $\hat{\underset{\sim}{\theta}}_n$, obtained by the Box-Cox maximum likelihood procedure, converges to $\underset{\sim}{\theta}_*' = (\mu_*,\sigma_*,\lambda_*)$ which is *the value of $\underset{\sim}{\theta}$ that minimizes the Kullback-Leibler information number* between the true density of the transformed variable f_λ and some *normal* p.d.f. $\phi_{\mu\sigma}$. Alternatively, under the Box-Cox Bayesian analysis, $\tilde{\lambda}_n$, the marginal posterior mode of λ converges to the same λ_*. These results provide a new meaning to the asymptotic value of the transformation

parameter. *The procedure seeks to minimize the mean information for discrimination between the densities of the transformed variable and a* normal, in the sense of the Kullback-Leibler information number. □

Remark

The known $g(\cdot)$ situation, corresponding to an unlimited sample size, represents the very best that can be done. Our theory thus provides bench marks for the maximum amount of improvement on approximate normality. However, it is clear that normality is not guaranteed by the Box-Cox procedure and that all the checks must be applied to the transformed data.

Our method, of selecting λ, has the advantage that we can measure numerically and graphically the improvement to normality introduced by the transformation. Numerically, we compare $I[g,\phi_{\mu_X\sigma_X}]$ with $I[f_{\lambda_*};\phi_{\mu_*,\sigma_*}]$ where $\mu_X = E_g(X)$ and $\sigma_X^2 = V_g(X)$. $I[g;\phi_{\mu_X,\sigma_X}]$ *represents* how 'far' the original p.d.f. $g(\cdot)$ is from a normal p.d.f. while $I[f_{\lambda_*};\phi_{\mu_*,\sigma_*}]$ measures how 'close' $f_{\lambda_*}(\cdot)$, the p.d.f. of the transformed variable, is to a normal p.d.f.

Examples

We compare both plots of $f_{\lambda_*}(\cdot)$ and $\phi_{\mu_*\sigma_*}(\cdot)$ and the information numbers $I[f_{\lambda_*};\phi_{\mu_*\sigma_*}]$, $I[g;\phi_{\mu_X\sigma_X}]$ for two families of p.d.f.'s. Our examples show how a power transformation can improve normality.

Example 1) The Gamma family. The p.d.f.'s are

$$g(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}e^{-\beta x} \quad , \quad \text{for} \quad x > 0; \quad \alpha,\beta > 0.$$

Since β is a scale parameter, it is sufficient to consider $\beta = 1$. Using the equalities

$$\Gamma^{(r)}(t) = \int_0^\infty [\ell og(u)]^r u^{t-1}e^{-u}du \quad , \quad r \geq 1$$

for derivatives, and

$$E_g(X^{\lambda r}) = \frac{\Gamma(\lambda r+\alpha)}{\beta^{\lambda r}\Gamma(\alpha)}, \quad \lambda r+\alpha > 0$$

it is readily shown that

$$G(\lambda) = \tfrac{1}{2}[\ell og(2\pi)+1]-2\ell og[\Gamma(\alpha)]+\alpha[\psi(\alpha)-1]-\lambda\psi(\alpha)$$
$$+ \tfrac{1}{2}\ell og\left\{\frac{\Gamma(\alpha)\Gamma(2\lambda+\alpha)-[\Gamma(\lambda+\alpha)]^2}{\lambda^2}\right\} \tag{3.8}$$

for $\lambda > -\frac{\alpha}{2}$. Here, ψ = digamma function, so $\psi(x) = \frac{d}{dx}\ell og[\Gamma(x)]$.

In Figure 1 we plot $G(\lambda)$ *vs.* λ *for* α *= 0.5, 1.0, 1.5, 2.0 and* 3.0. We can see that as α increases G becomes flatter around its minimum.

We note that although the selection of the transformation is invariant under scaling of X, the p.d.f. of $X^{(\lambda)}$, f_λ, is not. Figure 2 shows the improvement to normality for $\alpha = 1$ and $\beta = 1$, 5, 50 and Figure 3 presents plots of f_{λ_*} and $\phi_{\mu_*\sigma_*}$ for $\alpha = 2$ and β = .5, 5, 20. We can see that the improvement is quite remarkable. For small values of α, the effect of transforming is not that good. Figure 4 displays f_{λ_*} and $\phi_{\mu_*\sigma_*}$ for α = .5 and β = 1, 5, 10.

We finish the present example by noting that the 'normalizing' cube root transformation derived by Wilson and Hilferty (1931) is a "sensible choice. Using the information number approach we want to make $|G(\lambda_*)-G(1/3)|$ small as $\alpha \to \infty$. Employing the asymptotic formulas for $\Gamma(z+a)/\Gamma(z+b)$ and $\psi(z)$ (see Abramowitz and Stegun (1964) p. 257), (3.8) becomes

$$G(\lambda) \sim \text{const} + \frac{1}{4\alpha}(3\lambda^2-2\lambda+1) + (\alpha^{-2}), \quad \text{as} \quad \alpha \to \infty$$

uniformly for bounded sets of λ-values. Next, it is easily seen that λ = 1/3 minimizes (3.8) in the sense

$$G(\lambda)-G(1/3) = \frac{3(\lambda-1/3)^2}{4\alpha} + O(\alpha^{-2}).$$

Our plots of $G(\lambda)$ vs. λ show that λ = 1/3 is a sensible choice even for α as small as one. □

Example 2) The Inverse Gaussian family. The p.d.f. of the inverse Gaussian is

$$g(x) = \sqrt{\frac{\alpha}{2\pi}}\, x^{-3/2} \exp\{-\frac{\alpha(x-\mu)^2}{2\mu^2 x}\} \quad ; \text{ for } x > 0, \ \alpha,\mu > 0.$$

Let $K_\nu(z)$ *be the modified Bessel function of the second kind.* Using the equality (see Tweedie (1957), p. 363)

$$2(\tfrac{1}{2}z)^{-\nu}K_{\pm\nu}(z) = \int_0^\infty t^{-(1+\nu)}\exp\{-(t+\frac{z^2}{4t})\}dt, \tag{3.9}$$

where ν is not necessarily an integer, we obtain the fractional moments

$$E(X^{r\lambda}) = \sqrt{\frac{2\phi}{\pi}}\, e^{\phi}\mu^{r\lambda}K_{r\lambda-\frac{1}{2}}(\phi) \quad , \quad r \geq 1$$

and

$$V_g(X^{(\lambda)}) = \frac{\sqrt{2\phi}\, e^{\phi}\mu^{2\lambda}}{\lambda^2\sqrt{\pi}} \{K_{2\lambda-\frac{1}{2}}(\phi) - \sqrt{\frac{2\phi}{\pi}}\, e^{\phi}[K_{\lambda-\frac{1}{2}}(\phi)]^2\}$$

in terms of the ratio of parameters $\phi = \alpha/\mu$.

To determine $E_g[\ell og(X)]$, we differentiate (3.9) with respect to ν, to obtain

$$E_g[\ell og(X)] = \ell og(\mu) + Ei(-2\phi)e^{2\phi}$$

Here, $Ei(x) = ⨍_{-\infty}^{x} \frac{e^t}{t}dt$, where ⨍ denotes the Cauchy principal value of the integral. The function $G(\cdot)$ then becomes

$$G(\lambda) = \text{const}-\lambda Ei(-2\phi)e^{2\phi}+\tfrac{1}{2}\ell og\left\{\frac{1}{\lambda^2}\left[K_{2\lambda-\frac{1}{2}}(\phi)-\sqrt{\frac{2\phi}{\pi}}\,e^{\phi}[K_{\lambda-\frac{1}{2}}(\phi)]^2\right]\right\} \qquad (3.10)$$

where, const = $\frac{1}{2}\ell og(\phi)-\frac{1}{2}E_1(-2\phi)e^{2\phi}+\frac{1}{2}\ell og(\frac{2\phi}{\pi})+\phi$. Notice that (3.10) depends on α and μ only through ϕ. In Figure 5, we plot $G(\lambda)$ vs λ for ϕ = 2,3,4,5 and 10. We see that G becomes flatter around its minimum as ϕ increases. In Table (3.1), we record the values of λ_*, $I[f_{\lambda_*};\phi_{\mu_*\sigma_*}]$ and $I[g;\phi_{\mu_X\sigma_X}]$ for ϕ = 2, 3, 4 and 5. Also, in Figure 6, we plot f_{λ_*} and $\phi_{\mu_*\sigma_*}$ for $\phi = 2$. We see that the power transformation is achieving near normality.

Whitmore and Yalovsky (1978) propose the normalizing transformation

$$Y = \frac{1}{2\sqrt{\phi}} + \sqrt{\phi}\,\ell og(X)$$

corresponding to $\lambda = 0$. Note that $|G(0)-G(\lambda_*)|$ is small, even for ϕ as small as 2.

Employing the asymptotic expansion for $K_\nu(z)$ and $Ei(z) = -E_1(-z)$ (see Abramowitz and Stegun (1964), p. 378 and p. 231) it is easily shown that

$$G(\lambda) - G(0) = 3\lambda^2/4\phi + O(\phi^{-2}) \quad \text{as} \quad \phi \to \infty$$

uniformly on bounded sets of λ-values. Thus $\lambda = 0$ is a reasonable choice for large ϕ. □

Example 3 The Pareto distribution $g(x) = c^{-1}\alpha(x/\alpha)^{-\alpha}$, $x > c$ is graphed in Figure 7, for α=c=1, along with the pdf of the transformed variable $(X^{\lambda_*}-1)/\lambda_*$ with $\lambda_* = 2^{-\frac{1}{2}}$. The approximating normal is also shown. The power transformation is clearly unsuccessful in this situation □

The Table 3.1 allows us to measure numerically the improvement to normality introduced by the transformation. Note that the distributions are ordered according to the last column, that is, according to their flexibility of approaching the normal distribution by means of a power transformation.

Table 3.1

Comparison of Transformations and Information Numbers

Distribution	λ_*	$I[g;\phi_{\mu_X\sigma_X}]$	$I[f_{\lambda_*};\phi_{\mu_*\mu_*}]$
Gamma (3,β)	0.3124	0.12067	0.00019
Gamma (2,β)	0.3006	0.18830	0.00051
Inverse Gaussian φ = 5	-0.0256	0.13734	
Inverse Gaussian φ = 4	-0.0304	0.16842	0.00072
Inverse Gaussian φ = 3	-0.0363	0.21794	0.00111
Gamma (1.5,β)	0.2887	0.26070	0.00140
Inverse Gaussian φ = 2	-0.0502	0.30976	0.00198
Gamma (1,β) Neg. exponential	0.2654	0.41894	0.00278
Weibull (θ,β)	0.2654 × θ	depends on θ	
Gamma (0.5,β)	0.2084	0.93175	0.01205
Half Cauchy (θ)	0.0000	+ ∞	0.03264
Pareto (α,c)	$-\frac{\alpha}{\sqrt{2}}$	depends on α	0.15056

Table (3.1) suggests that all the probability distributions having p.d.f.'s may be ordered according to their Kullback-Leibler information number with their 'best' approximating normal distribution, that is to say, according to $I[g;\phi_{\mu_x,\sigma_x}]$.

4. Transformations of Multivariate Observations and to Linear Models

Transformation of Multivariate Observations

Let $\underset{\sim}{X} = (X_1,\ldots,X_p)'$ be a p-dimensional positive random vector, that is, $X_i > 0$ with probability one for $i = 1,\ldots,p$. Let $g(\cdot)$ be the p.d.f. of $\underset{\sim}{X}$, which is assumed to be known, and let $\underset{\sim}{Y}$ be the vector of transformed random variables, $\underset{\sim}{Y} = (X_1^{(\lambda_1)},\ldots,X_p^{(\lambda_p)})'$, where $X_i^{(\lambda_i)}$ is defined as in (1.1). *Further*, let $\phi_{\underset{\sim}{\mu}\Sigma}$ be the p.d.f. of a p-variate normal distribution with mean $\underset{\sim}{\mu}$ and positive definite covariance matrix $\Sigma = (\sigma_{ij})$.

Proposed Procedure

In order to determine $\underset{\sim}{\lambda} = (\lambda_1,\ldots,\lambda_p)'$ so that $\underset{\sim}{Y}$ has an approximate p-variate normal distribution, minimize the Kullback-Leibler information number $I[f_{\underset{\sim}{\lambda}};\phi_{\underset{\sim}{\mu}\Sigma}]$, between f_λ and $\phi_{\underset{\sim}{\mu}\Sigma}$, with respect to $\underset{\sim}{\mu}$, Σ and $\underset{\sim}{\lambda}$. □

We notice that, in the present situation, we have the flexibility of choosing the covariance structure of the transformed vector $\underset{\sim}{Y}$. *For instance, we may want to find* $\underset{\sim}{\lambda}$ *so that the covariance matrix* of $\underset{\sim}{Y}$ is diagonal. In this case we minimize $I[f_{\underset{\sim}{\lambda}};\phi_{\underset{\sim}{\mu}\Sigma}]$ subject to $\Sigma = \mathrm{diag}(\sigma_{11},\ldots,\sigma_{pp})$. We will develop our results for a matrix Σ with no particular structure.

Lemma 4.2 determines $\underset{\sim}{\mu}$ and Σ so that $I[f_{\underset{\sim}{\lambda}};\phi_{\underset{\sim}{\mu}\Sigma}]$ is minimum for fixed $\underset{\sim}{\lambda}$. That proof is based on the following well known result (see Watson (1964)).

<u>Lemma 4.1</u> Let A and B be $p \times p$ symmetric positive definite matrices. Let $\psi(A) = \mathrm{tr}(AB) - \ell og[\det(A)]$; $\mathrm{tr}(AB)$ = trace of the matrix AB and $\det(A)$ stands for the determinant of A. Then ψ achieves its minimum at $A = B^{-1}$. □

<u>Lemma 4.2</u> Let $\underset{\sim}{\lambda}$ be fixed. Assume $E_g(X_i^{\lambda_i}X_j^{\lambda_j})$, $E_g[X_i^{\lambda_i}\ell og(X_j)]$ and $E_g[\ell og(X_i)\ell og(X_j)]$ are finite for all $i,j = 1,\ldots,p$. Then, the values of $\underset{\sim}{\mu}$ and Σ that minimize $I[f_{\underset{\sim}{\lambda}};\phi_{\underset{\sim}{\mu}\Sigma}]$ are

$$\underset{\sim}{\mu}_*(\underset{\sim}{\lambda}) = \left[E_g(X_1^{(\lambda_1)}),\ldots,E_g(X_p^{(\lambda_j)})\right]'$$

and

$$\Sigma_*(\lambda) = \left\{E_g\left[X_i^{(\lambda_i)} - E_g(X_i^{(\lambda_j)})\right]\left[X_j^{(\lambda_j)} - E_g(X_j^{(\lambda_j)})\right]\right\} \; ; \quad i,j = 1,\ldots,p.$$

That is, the first and second order moments of the normal distribution match those of the transformed variable.

<u>Proof</u> Under the moment assumptions we have the finite information number

$$I[f_{\underset{\sim}{\lambda}};\phi_{\underset{\sim}{\mu}\Sigma}] = E_g\{\ell og\ g(\underset{\sim}{X})\}+\sum_{i=1}^{p}(1-\lambda_i)E_g\{\ell og(X_i)\}+\tfrac{1}{2}\ell og[(2\pi)^p\det(\Sigma)]$$

$$+\tfrac{1}{2}tr\{\Sigma^{-1}C\}+\tfrac{1}{2}tr\{\Sigma^{-1}B\}$$

where $C = E_g\{[\underset{\sim}{Y}-E_g(\underset{\sim}{Y})][\underset{\sim}{Y}-E_g(\underset{\sim}{Y})]'\}$ *and* $B = (E_g(\underset{\sim}{Y})-\mu)(E_g(\underset{\sim}{Y})-\mu)'$.

Also

$$H(\underset{\sim}{\lambda})=\min_{\underset{\sim}{\mu},\Sigma} I[f_{\underset{\sim}{\lambda}};\phi_{\underset{\sim}{\mu}\Sigma}] = E_g\{\ell og[g(\underset{\sim}{X})]\}+\sum_{i=1}^{p}(1-\lambda_i)E_g\{\ell og(X_i)\}+\frac{p}{2}\ell og(2\pi)$$

$$+\tfrac{1}{2}\min_{\mu,\Sigma}\{\ell og[\det(\Sigma)]+tr(\Sigma^{-1}C)+tr(\Sigma^{-1}B)\}\ . \qquad (4.1)$$

$$\geq \frac{p}{2}[\ell og(2\pi)+1]+E_g\{\ell og[g(\underset{\sim}{X})]\} + \sum_{i=1}^{p}(1-\lambda_i)E_g[\ell og(X_i)]+\tfrac{1}{2}\ell og[\det(C)] \qquad (4.2)$$

by Lemma 4.1 and because $tr(\Sigma^{-1}B) \geq 0$. The lower bound (4.2) is achieved by $\underset{\sim}{\mu}_*(\underset{\sim}{\lambda})$ and $\Sigma_*(\lambda)$. □

For the specified choices $\underset{\sim}{\mu}_*(\lambda)$ and $\Sigma_*(\lambda)$, (4.1) reduces to

$$H(\underset{\sim}{\lambda}) = \frac{p}{2}[\ell og(2\pi)+1]+E_g\{\ell og[g(\underset{\sim}{X})]\}+\sum_{i=1}^{p}(1-\lambda_i)E_g[\ell og(X_i)]$$

$$+\tfrac{1}{2}\ell og\{\det[\Sigma_*(\lambda)]\}\ . \qquad (4.3)$$

The optimal value, $\underset{\sim}{\lambda}_*$, *is found by minimizing* (4.3). *It is clear from* (4.3) that the selection of $\underset{\sim}{\lambda}_*$ is scale invariant.

Writing (4.3) in terms of the correlation matrix, ρ, of the transformed vector $\underset{\sim}{Y}$ we can appreciate some interesting properties of our method of selecting $\underset{\sim}{\lambda}_*$. The matrix $\rho = (\rho_{ij})$ is defined as

$$\rho_{ij} = \frac{E_g[X_i^{(\lambda_i)}X_j^{(\lambda_j)}]-E_g[X_i^{(\lambda_i)}]E_g[X_j^{(\lambda_j)}]}{\sqrt{V_g(X_i^{(\lambda_i)})V_g(X_j^{(\lambda_j)})}}$$

and it is well known that $\rho = D^{-1}\Sigma_*(\underset{\sim}{\lambda})D^{-1}$ where

where $D = \text{diag}\ V_{g_1}^{\frac{1}{2}}(X_1^{(\lambda_1)}),\ldots,V_{g_p}^{\frac{1}{2}}(X_p^{(\lambda_p)})$ and $\det(\rho) > 0$

whenever $\Sigma_*(\underset{\sim}{\lambda})$ is non-singular. Thus

$$H(\underset{\sim}{\lambda}) = \sum_{i=1}^{p} G_i(\lambda_i)+\tfrac{1}{2}\ell og[\det(\rho)]+E_g\{\ell og\ g(\underset{\sim}{X})\}-\sum_{i=1}^{p}E_{g_i}\{\ell og\ g_i(X_i)\} \qquad (4.4)$$

where g_i is the marginal p.d.f. of X_i and G_i is the function G defined in (3.7) for the random variable X_i.

When the random $\underset{\sim}{X}$ has independent components, that is, $g(x_1,\ldots,x_p) = g_1(x_1)\ \ldots\ g_p(x_p)$ then $\rho = I_p$ and hence (4.4) becomes $H_I(\underset{\sim}{\lambda}) = \Sigma\ G_i(\lambda_i)$ where we use the subscript I on H to denote the condition that the components are independent. Thus selecting $\underset{\sim}{\lambda}$ under independence (or zero correlation) is equivalent to the separate selection λ_i according to the marginal distribution.

On the other hand, when the approximating normal has covariance $\Sigma = \text{diag}(\sigma_{11},\ldots,\sigma_{pp})$, (4.4) becomes

$$H_1(\underset{\sim}{\lambda}) = \sum_{i=1}^{p} G_i(\lambda_i)+E_g\{\ell og\ g(\underset{\sim}{X})\}-\sum_{i=1}^{p}E_{g_i}\{\ell og\ g_i(X_i)\} = H_I(\underset{\sim}{\lambda})+\text{constant}.$$

Therefore, when Σ is a diagonal matrix the selection of $\underset{\sim}{\lambda}$ is again equivalent to finding $\underset{\sim}{\lambda}$ when the components of the original vector $\underset{\sim}{X}$ are independent. Further restricting Σ to $\sigma^2 I_p$, (4.4) becomes

$$H_2(\underset{\sim}{\lambda}) = \frac{p}{2}[\log(2\pi)+]+E_g\{\log g(\underset{\sim}{X})\}+\sum_{i=1}^{p}(1-\lambda_i)E_{g_i}\{\log X_i\}+\frac{p}{2}\log\left\{\frac{1}{p}\sum_{i=1}^{p}V_{g_i}(X_i^{(\lambda_i)})\right\} .$$

It is easy to show that $H(\underset{\sim}{\lambda}) \le H_1(\underset{\sim}{\lambda}) \le H_2(\underset{\sim}{\lambda})$ which reflects the general fact that the more we restrict Σ the larger the value of $\min_{\underset{\sim}{\mu},\Sigma,\underset{\sim}{\lambda}} I[f_{\underset{\sim}{\lambda}};\phi_{\underset{\sim}{\mu}\Sigma}]$.

<u>Example 1</u> We chose the bivariate gamma distribution type I (See Johnson and Kotz (1972)) with p.d.f.

$$g_{X_1X_2}(x_1,x_2) = e^{-x_1-x_2}(e^{\min(x_1,x_2)}-1) \quad \text{for} \quad x_1,x_2 > 0.$$

Since X_i is gamma with shape parameter = 2 and scale parameter = 1 we know, from example 1 of section 3 that $V_{g_i}(X_i^{(\lambda_i)}) = \{\Gamma(2+2\lambda_i)-[\Gamma(2+\lambda_i)]^2\}/\lambda_i^2$; for $\lambda_i > -\frac{1}{2}$; $i = 1,2$. After some algebra it is seen that

$$E_g(X_1^{\lambda_1}X_2^{\lambda_2}) = \frac{\Gamma(\lambda_1+\lambda_2+3) - \Gamma(2+\lambda_1)\Gamma(2+\lambda_2)}{(1+\lambda_1)(1+\lambda_2)}$$

for $\lambda_1+\lambda_2+3 > 0$, and hence the function H in (4.3) becomes

$$H(\lambda_1,\lambda_2) = \log(2\pi)+1+E_g\{\log[g(\underset{\sim}{X})]\}+(1-\gamma)(2-\lambda_1-\lambda_2)$$
$$+ \log\{\det[\Sigma_*(\lambda_1,\lambda_2)]\} \quad . \qquad (4.5)$$

We notice that $H(\lambda_1,\lambda_2) = H(\lambda_2,\lambda_1)$ so that the two components of the

vector $\underset{\sim}{\lambda}_*$ are equal. Here $\underset{\sim}{\lambda}_* = (0.3237, 0.3237)'$. From Table 3.1 we see that the separate selection of each λ_{*i} gives $\lambda_{*s} = (0.3006, 0.3006)'$ where the subscript s is used to denote that each component was found separately. If Σ were assumed to be diagonal, these are the values we would have obtained.

The 'best' approximating normal distribution has mean $\mu_1(\lambda_*) = \mu_2(\lambda_*) = 1.1836$ and $\sigma_{11}(\lambda_*) = \sigma_{22}(\lambda_*) = .7759$ and $\sigma_{12}(\lambda_*) = .3727$.

Some contours of f_{λ_*} and $\phi_{\underset{\sim}{\mu}_*\Sigma_*}$ are plotted in Figure 8. □

<u>Remark</u> Generally speaking, we would expect the joint normality to be most improved by fitting a general $\underset{\sim}{\mu}$ and Σ. Considering only marginal transformations, in order to reduce the computations, is now seen to be equivalent to forcing the approximating normal to have independent components. A situation like this could also be called transformation to independence. □

Again given enough regularity conditions, the population procedure can be shown to be the large sample limit of that based on maximizing

$$\ell(\underset{\sim}{\mu},\Sigma,\underset{\sim}{\lambda}|\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n) = -\frac{np}{2}\log(2\pi)-\frac{n}{2}\log[\det(\Sigma)] + \sum_{j=1}^{n}\sum_{i=1}^{p}(\lambda_i-1)\log(x_{ij})$$
$$-\tfrac{1}{2}\mathrm{tr}[\Sigma^{-1}\sum_{j=1}^{n}(\underset{\sim}{x}_j^{(\underset{\sim}{\lambda})}-\underset{\sim}{\mu})(\underset{\sim}{x}_j^{(\underset{\sim}{\lambda})}-\underset{\sim}{\mu})']$$

over $\underset{\sim}{\mu},\Sigma$ and $\underset{\sim}{\lambda}$. Here $\underset{\sim}{x}_j^{(\underset{\sim}{\lambda})} = (x_{1j}^{(\lambda_1)},\ldots,x_{pj}^{(\lambda_p)})'$.

<u>Transformation to a linear Model</u>

Let $\underset{\sim}{Y} = (Y_1,\ldots,Y_n)$ be an n-dimensional positive random vector with known p.d.f. $g(\cdot)$ and marginal p.d.f.'s $g_i(\cdot)$ of Y_i. Furthermore, let $\underset{\sim}{U} = \underset{\sim}{Y}^{(\lambda)} = (Y_1^{(\lambda)},\dots,Y_n^{(\lambda)})'$ have p.d.f. $f_\lambda(\cdot)$ and $\phi_{X\beta,\sigma^2 I_n}(\cdot)$ be the p.d.f. of a normal distribution with mean $X\underset{\sim}{\beta}$ and covariance matrix $\sigma^2 I_n$. We are interested in transforming $\underset{\sim}{Y}$ in such a way that *the distribution of the transformed vector* $\underset{\sim}{U}$ *more closely approximate* a normal linear model with a given $n\times p$ design matrix X, of rank p.

Proposed Procedure

Select a transformation λ, to minimize the Kullback-Leibler information number $I[f_\lambda;\phi_{X\underset{\sim}{\beta},\sigma^2 I_n}]$ with respect to $\underset{\sim}{\beta},\sigma$ and λ, simultaneously. □

We minimize the information number sequentially. That is, we first find the values of $\underset{\sim}{\beta}$ and σ^2 that best approximate $\phi_{X\underset{\sim}{\beta},\sigma^2 I_n}$ by f_λ and then search for the minimizing value of λ. Lemma 4.3 determines for fixed λ, the values of $\underset{\sim}{\beta}$ and σ^2 that make $f_\lambda(\cdot)$ 'closest' to $\phi_{X\underset{\sim}{\beta},\sigma^2 I_n}(\cdot)$ and Lemma 4.4 shows that the selection of λ is scale invariant, provided that the design matrix X has a column of ones.

Lemma 4.3 Let λ be fixed. Assume that $E_g(Y_i^\lambda Y_j^\lambda)$, $E_g[Y_i^\lambda \log(Y_j)]$ and $E_g[\log(Y_i)\log(Y_j)]$ are finite for all $i,j = 1,\dots,n$. Then, the values of $\underset{\sim}{\beta}$ and σ^2 that minimize $I[f_\lambda;\phi_{X\underset{\sim}{\beta},\sigma^2 I_n}]$ are

$$\underset{\sim}{\beta}_*(\lambda) = (X'X)^{-1}X'E_g[\underset{\sim}{Y}^{(\lambda)}]$$

and

$$\sigma_*^2(\lambda) = \frac{1}{n}E_g[\underset{\sim}{Y}^{(\lambda)\prime}(I_n-P)\underset{\sim}{Y}^{(\lambda)}]$$

where $P = X(X'X)^{-1}X'$.

Proof A proof is given in Hernández (1978).

Let $G(\lambda) = \min_{\underset{\sim}{\beta},\sigma^2} I[f_\lambda;\phi_{X\underset{\sim}{\beta},\sigma^2 I_n}]$. Then according to Lemma 4.3

$$G(\lambda) = \frac{n}{2}[\log(2\pi)+1]+E_g\{\log[g(\underset{\sim}{Y})]\}+(1-\lambda)\sum_{i=1}^{n} E_{g_i}[\log(Y_i)]+\frac{n}{2}\log[c_*^2(\lambda)]. \quad (4.6)$$

Lemma 4.4 Let $\alpha > 0$ and set $\underset{\sim}{Y}_* = \alpha\underset{\sim}{Y}$ and $\underset{\sim}{Y}_*^{(\lambda)} = (\alpha\underset{\sim}{Y})^{(\lambda)}$. Let $g_*(\cdot)$ *and* $f_\lambda^*(\cdot)$ *be the p.d.f.'s of* $\underset{\sim}{Y}_*$ *and* $Y_*^{(\lambda)}$, *respectively. Then*

$$G_*(\lambda) = \min_{\underset{\sim}{\beta},\sigma^2} I[f_\lambda^*;\phi_{X\underset{\sim}{\beta},\sigma^2 I_n}] = G(\lambda)$$

where $G(\lambda)$ is given in (4.6) provided that the matrix X contains a column consisting of ones.

Proof A proof is given in Hernández (1978).

We now describe how one could numerically evaluate the relative contribution of each of the 'ideal conditions' of least squares to the final selection of λ_*. This breakdown of the criterion into components is the population analogue of the sample decomposition proposed by Box and Cox (1964).

Let $G(\lambda|N,H,S)$ denote the function when we want a power transformation to achieve, *Normality (N), Homogeneity of variances (H) and Simplicity* of structure in the expections (S). The decomposition

$$\min_\lambda G(\lambda|N,H,S) = \min_\lambda G(\lambda|N)+\{\min_\lambda G(\lambda|N,H)-\min_\lambda G(\lambda|N)\} \quad (4.7)$$

$$+ \{\min_\lambda G(\lambda|N,H,S)-\min_\lambda G(\lambda|N,H)\}$$

partitions the overall criterion into three parts

1) $\min_{\lambda} G(\lambda|N)$, where $G(\lambda|N) = \min_{\underset{\sim}{\mu},\Sigma} I[f_{\lambda};\phi_{\underset{\sim}{\mu}\Sigma}]$, measures the contribution of the transformation to normality.

2) $\{\min_{\lambda} G(\lambda|N,H) - \min_{\lambda} G(\lambda|N)\}$, is the contribution of the additional requirement of homogeneity of variances, given that normality has been already incorporated. Here, $G(\lambda|N,H) = \min_{\underset{\sim}{\mu},\sigma^2} I[f_{\lambda};\phi_{\underset{\sim}{\mu},\sigma^2 I_n}]$, where $\underset{\sim}{\mu}$ is restricted to be of the form $X_1\underset{\sim}{\gamma}$ representing a 'larger' model (e.g. including interactions).

3) $\{\min_{\lambda} G(\lambda|N,H,S) - \min_{\lambda} G(\lambda|N,H)$ is the contribution of simplicity of structure in the expectations (e.g. we want additivity), given that Normality and homogeneity of variances have been included.

Remark It is worth noting that, if the original random vector $\underset{\sim}{Y}$ does not have independent components, the contribution of homogeneity of variances is confounded with that of the requirement of independence. □

Simultaneous plots of $G(\lambda|N)$, $\{G(\lambda|N,H)-G(\lambda|N)\}$, $\{G(\lambda|N,H,S)-G(\lambda|N,H)\}$ and $G(\lambda,N,H,S)$ vs. λ would display graphically the contributions of the different requirements imposed on the model. The above analysis is a parallel of Box and Cox (1964), pp. 226-37, and under appropriate regularity conditions, it is tne 'infinite' sample version.

APPENDIX

In this appendix we give the proofs of Theorems 2.2 and 2.3.

Proof of Theorem 2.2. To establish our results we employ a theorem *on uniform convergence* (Rubin (1956)). We now check the conditions of *this theorem.* Let $\ell:\Theta\times(0,\infty) \to R$ be defined as

$$\ell(\underset{\sim}{\theta}|x) = -\tfrac{1}{2}\ell og(2\pi) - \ell og(\sigma) + (\lambda-1)\ell og(x) - \frac{1}{2\sigma^2}(x^{(\lambda)}-\mu)^2. \qquad (A.1)$$

It is easy to verify that $|\ell(\underset{\sim}{\theta}|x)| \leq h(x)$ for all $\theta\in\Theta$ and $x > 0$ where

$$h(x) = \tfrac{1}{2}\ell og(2\pi) + |\ell og(c)\ | + |\ \ell og(d)\ | + (L+1)|\ell og(x)|$$

$$+ \frac{1}{c^2}\{[x^{(a)}]^2 + [x^{(b)}]^2 + M^2\}$$

and $L = \max\{|a|,b\}$. Since $[\ell og(x)]^2 \leq [x^{(a)}]^2 + [x^{(b)}]^2$, assumption *ii) guarantees that* $E_g[h(X)] < \infty$. Also, letting $S_i = [\frac{1}{i},i]$, $i \geq 1$ so that, $(0,\infty) - \bigcup_{i=1}^{\infty} S_i$ = the *empty set, which always has probability zero.*

Finally, ℓ, as defined in (A.1) is continuous in (θ',x). Since $\Theta \times S_i$ is compact, it follows that ℓ is uniformly continuous in (θ',x). Let $\| \ \|_p$ denote the Euclidean norm in the p-space then, for any $\epsilon > 0$, $|\ell(\theta|x)-\ell(\theta_1|x_1)| < \epsilon$ whenever $\|(\theta',x)-(\theta_1',x_1)\|_4 < \delta(\epsilon)$. Thus, by setting $x_1 = x$ we obtain $|\ell(\theta|x)-\ell(\theta_1|x)| < \epsilon$ whenever $\|\theta-\theta_1\|_3 < \delta(\epsilon)$ for all $x \in S_i$. That is, $\ell(\theta|x)$ is equicontinuous in θ for $x \in S_i$.

We conclude that, with probability one,

$$\lim_{n\to\infty} \frac{1}{n}\ell(\theta|X_n) = E_g[\ell(\theta|X)] \qquad (A.2)$$

uniformly in θ for $\theta \in \Theta$ and that the limit function is continuous in θ. Equivalently

$$\lim_{n\to\infty}\{\max_{\theta\in\Theta}|\frac{1}{n}\ell(\theta|X_n)-E_g[\ell(\theta|X)]|\} = 0 \text{ , with probability one} \qquad (A.3)$$

The result 1) follows directly since

$$|\max_{\theta\in\Theta}\{\frac{1}{n}\ell(\theta|X_n)\}-\max_{\theta\in\Theta}\{E_g[\ell(\theta|X)]\}| \le \max_{\theta\in\Theta}|\frac{1}{n}\ell(\theta|X_n)-E_g[\ell(\theta|X)]|$$

which tends to zero by (A.3).

2) Since we are establishing almost sure convergence, we introduce the notation ω for a generic outcome and A for the set where (A.3) holds. To obtain a contradiction, we assume that $\hat{\theta}_n(\omega) \xrightarrow{a.s.}\!\!\!\!\!\!/ \ \theta_0$, so there exists a set of outcomes B where

$\hat{\theta}_n(\omega) \not\to \theta_0$ and $P(B) > 0$. We restrict our attention to the set $C = A \cap B$ with $P(C) > 0$.

Since Θ is compact, for each $\omega \in C$ there exists a subsequence $\{m\} \subset \{n\}$ and a limit point $\theta_*(\omega)$ with $\hat{\theta}_m(\omega) \to \theta_*(\omega) \neq \theta_0$. However, by definition of $\hat{\theta}_m$,

$$\frac{1}{m}\ell(\theta_0|X_m) \le \frac{1}{m}\ell(\hat{\theta}_m|X_m) \qquad \text{for each} \quad \omega \in C \text{ .} \qquad (A.4)$$

Also

$$\begin{aligned} |\tfrac{1}{m}\ell(\hat{\theta}_m|X_m)-E_g[\ell(\theta_*|X)]| &\le |\tfrac{1}{m}\ell(\hat{\theta}_m|X_m)-E_g[\ell(\hat{\theta}_m|X)]|+|E_g[\ell(\hat{\theta}_m|X)] \\ &\qquad -E_g[\ell(\theta_*|X)]| \\ &\le \max_{\theta\in\Theta}|\tfrac{1}{m}\ell(\theta|X_m)-E_g[\ell(\theta|X)]| \\ &\qquad +|E_g[\ell(\hat{\theta}_m|X)]-E_g[\ell(\theta_*|X)]| \text{ .} \end{aligned} \qquad (A.5)$$

For $\omega \in C$, we take the limit as $m \to \infty$ on the right hand side of (A.5). The first term goes to zero by (A.3) and the second also goes to zero by the continuity of $E_g[\ell(\cdot|X)]$ as stated after (A.2) and the fact that $\hat{\theta}_m(\omega) \to \theta_*(\omega)$ on C. Returning to (A.4) and taking the limit as $m \to \infty$, we obtain

$$E_g[\ell(\theta_0|X)] \le E_g[\ell(\theta_*|X)] \quad \text{, each} \quad \omega \in C,$$

which is a contradiction to the assumption iii) which states that θ_0 is the unique global maximum. Thus $\hat{\theta}_n \to \theta_0$, with probability one.

3) To establish the asymptotic normality of the M.L.E. $\hat{\theta}_n$, we expand the product $n^{-\frac{1}{2}}$ times the gradient of the log-likelihood function

$$\frac{1}{\sqrt{n}}\nabla\ell(\hat{\theta}_n|X_n) = \frac{1}{\sqrt{n}}\nabla\ell(\theta_0|X_n)+\frac{1}{n}\nabla^2\ell(\theta_{*n}|X_n)\sqrt{n}(\hat{\theta}_n-\theta_0)$$

where $\theta_* = \alpha_n\hat{\theta}_n+(1-\alpha_n)\theta_0$ with $\alpha_n\in(0,1)$, $n \geq 1$.

By assumption iv), θ_0 is an interior point of Θ and since $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$ the derivatives vanish at the maximum and $\nabla\ell(\hat{\theta}_n|X_n) \to 0$ with probability one. Consequently,

$$\frac{1}{\sqrt{n}}\nabla\ell(\theta_0|X_n)-[-\frac{1}{n}\nabla^2\ell(\theta_{*n}|X_n)]\sqrt{n}(\hat{\theta}_n-\theta_0) \to 0 \qquad (A.6)$$

in probability. Left multiplying (A.6) by the matrix $V = \{-E_g[\nabla^2\ell(\theta_0|X)]\}^{-1}$, which exists according to assumption vii), does not change the convergence in probability to 0. However, applying the Central Limit Theorem to $n^{-\frac{1}{2}}\nabla\ell(\theta_0|X_n)$ and employing assumption vi)

$$V\frac{1}{\sqrt{n}}\nabla\ell(\theta_0|X_n) \xrightarrow{d} N(0,VWV')$$

where $W = E_g\{\nabla\ell(\theta_0|X)[\nabla\ell(\theta_0|X)]'\}$, thus

$$V[-\frac{1}{n}\nabla^2\ell(\theta_{*n}|X_n)]\sqrt{n}(\hat{\theta}_n-\theta_0) \xrightarrow{d} N(0,VWV') .$$

Next, according to Lemma 2.1, $\frac{\partial^2\ell(\theta|x)}{\partial\theta_i\partial\theta_j}$ is continuous in (θ,x) for each pair (i,j) and is dominated in absolute value by a function which is g-integrable according to assumption v)

Thus, with $S_i = [\frac{1}{i},i]$ we have that $\frac{\partial^2\ell(\theta|x)}{\partial\theta_i\partial\theta_j}$ is equicontinuous in θ for $x\in S_i$. Hence, applying Rubin (1956) we conclude that

$$\lim_{n\to\infty}\frac{1}{n}\frac{\partial^2\ell(\theta|X_n)}{\partial\theta_i\partial\theta_j} = E_g\left[\frac{\partial^2\ell(\theta|X)}{\partial\theta_i\partial\theta_j}\right] \text{ with probability one } (A.7)$$

and this convergence is uniform in θ. Moreover, the limit function is continuous in θ. Equivalently,

$$\lim_{n\to\infty}\max_{\theta\in\Theta}\|\frac{1}{n}\nabla^2\ell(\theta|X_n)-E_g[\nabla^2\ell(\theta|X)]\|_g = 0 . \qquad (A.8)$$

Next, consider the difference

$$\|\frac{-1}{n}\nabla^2\ell(\theta_{*n}|X_n)-V^{-1}\|_g \leq \|\frac{-1}{n}\nabla^2\ell(\theta_{*n}|X_n)-E_g[-\nabla^2\ell(\theta_{*n}|X)]\|_g$$
$$+\|E_g[-\nabla^2\ell(\theta_{*n}|X)]-V^{-1}\|_g$$
$$\leq \max_{\theta\in\Theta}\|\frac{-1}{n}\nabla^2\ell(\theta|X_n)-E_g[-\nabla^2\ell(\theta|X)]\|_g$$
$$+\|E_g[-\nabla^2\ell(\theta_{*n}|X)]-V^{-1}\|_g . \qquad (A.9)$$

For each outcome ω in the set where (A.8) holds, we take the limit as $n \to \infty$ on the right hand side of (A.9). The first term tends to zero by (A.8) and the second also tends to zero by the continuity of $E_g[\nabla^2 \ell(\cdot|X)]$ established after (A.7), and the fact that $\hat{\underline{\theta}}_n(\omega) \to \underline{\theta}_0$. Hence, $-\frac{1}{n}\nabla^2\ell(\underline{\theta}_{*n}|\underline{X}_n) \xrightarrow{a.s.} V^{-1}$. Thus, $\{I_3 - V[-\frac{1}{n}\nabla^2\ell(\underline{\theta}_{*n}|\underline{X}_n)]\}\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_0) \to \underline{0}$ in probability and by Slutsky's Theorem we conclude that

$$\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_0) \xrightarrow{d} N(\underline{0}, VWV') . \quad \square$$

<u>Proof of Theorem 2.3.</u> From (1.6) we obtain

$$\frac{1}{n}\ell og[p(\lambda|\underline{X}_n)] - \frac{1}{n}\ell og(C_n) = \frac{\nu}{n}\left\{(\lambda-1)\frac{\sum_{i=1}^{n}\ell og(X_i)}{n} - \tfrac{1}{2}\ell og[S^2(\lambda)]\right\} + \frac{1}{n}\ell og[p(\lambda)]$$

or

$$\frac{1}{n}\ell og[p(\lambda|\underline{X}_n)] - \frac{1}{n}\ell og(C_n) = \frac{\nu}{n}\left\{\frac{1}{n}\ell_{max}(\lambda) + \tfrac{1}{2}[\ell og(2\pi)+1] + \tfrac{1}{2}\ell og(\tfrac{\nu}{n})\right\} + \frac{1}{n}\ell og[p(\lambda)] \qquad (A.11)$$

where $C_n = C_n(\underline{X}_n)$ is the normalization constant and $\ell_{max}(\lambda)$ is defined as

$$\ell\ max(\lambda) = \max_{\mu,\sigma} \ell(\underline{\theta}|\underline{x}_n) = \frac{n}{2}[1+\ell og(2\pi)] - \frac{n}{2}\ell og[\frac{\nu}{n}S^2(\lambda)] + (\lambda-1)\sum_{i=1}^{n}\ell og(x_i).$$

Since we are interested in the posterior mode $\tilde{\lambda}_n$, in the following we can neglect the term $\frac{1}{n}\ell og(C_n)$ which does not depend on λ.

Moreover,

$$|\max_{\mu,\sigma}\frac{1}{n}\ell(\underline{\theta}|\underline{X}_n) - \max_{\mu,\sigma}E_g[\ell(\underline{\theta}|X)]| \le \max_{\mu,\sigma}|\frac{1}{n}\ell(\underline{\theta}|\underline{X}_n) - E_g[\ell(\underline{\theta}|X)]|$$

$$\le \max_{\mu,\sigma,\lambda}|\frac{1}{n}\ell(\underline{\theta}|\underline{X}_n) - E_g[\ell(\underline{\theta}|X)]|$$

which, according to (A.3) converges to zero with probability one. Since $\ell_{max}(\lambda)/n = \max_{\mu,\sigma}\ell(\underline{\theta}|\underline{X}_n)/n$ we conclude that

$$\lim_{n\to\infty}\{\frac{1}{n}\ell_{max}(\lambda)\} = \max_{\mu,\sigma}E_g[\ell(\underline{\theta}|X)], \text{ with probability one} \qquad (A.12)$$

and the convergence is uniform in λ. Also, by assumption, $p(\lambda)$ is continuous and positive on $[a,b]$ so that $|\ell og[p(\lambda)]|$ is bounded for $\lambda \in [a,b]$. Hence, $\lim_{n\to\infty}\{\frac{1}{n}\ell og[p(\lambda)]\} = 0$ uniformly in λ. Returning to (A.10) and taking the limit as $n \to \infty$ we have, using (A.11) and (A.12), that

$$\lim_{n\to\infty}\left\{\frac{\nu}{n}\left\{(\lambda-1)\frac{\sum_{i=1}^{n}\ell og(X_i)}{n} - \tfrac{1}{2}\ell og[S^2(\lambda)]\right\} + \frac{1}{n}\ell og[p(\lambda)]\right\}$$

$$= \max_{\mu,\sigma}E_g[\ell(\underline{\theta}|X)] + \tfrac{1}{2}[\ell og(2\pi)+1], \qquad (A.13)$$

with probability one, and this convergence is uniform in λ. Thus,

$$\lim_{n\to\infty}\{\frac{1}{n}\ell og[p(\lambda|\underline{X}_n)] - \frac{1}{n}\ell og(C_n)\} = \max_{\mu,\sigma}E_g[\ell(\underline{\theta}|X)] + \tfrac{1}{2}[\ell og(2\pi)+1], \qquad (A.14)$$

with probability one, uniformly in λ. Moreover,

$$\max_{\lambda}\{\frac{1}{n}\ell_{max}(\lambda)\} = \max_{\mu,\sigma,\lambda}\{\frac{1}{n}\ell(\underline{\theta}|\underline{X}_n)\} \to \max_{\mu,\sigma,\lambda}E_g[\ell(\underline{\theta}|X)] \text{ with probability one}$$

and the maximum is reached, by assumption, at $\underline{\theta}_0$. As a consequence of (A.11), the uniform convergence in (A.14) and a proof by contradiction, similar to that in part 2) of the proof of Theorem 2.2 we have that $\tilde{\lambda}_n \to \lambda_0$ with probability one. $\square$

Function G(·) for the Gamma Distribution

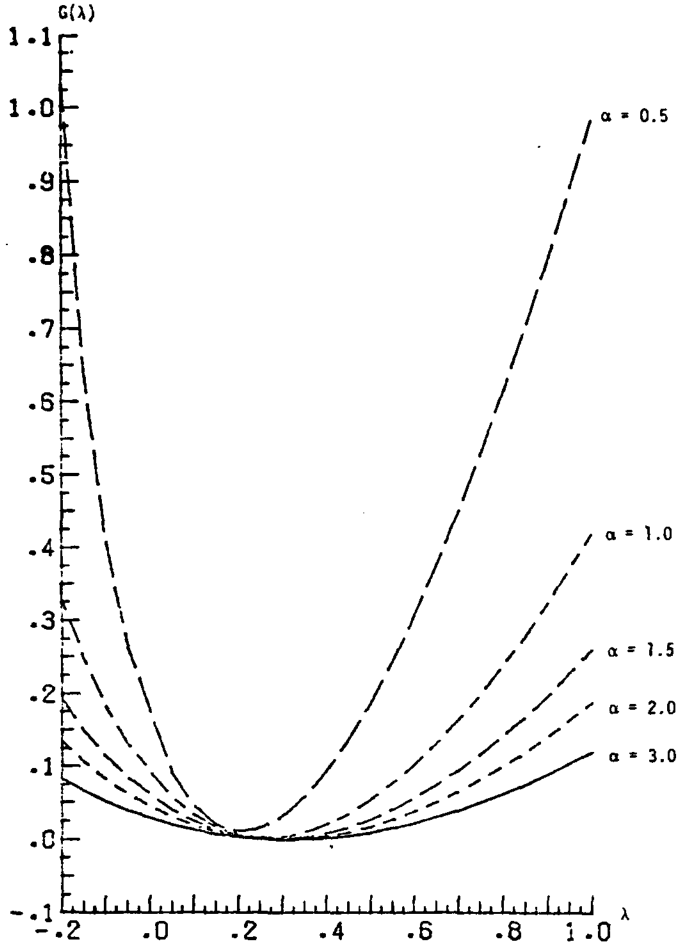


Figure 1

Plots of f_{λ_*} (solid lines) and $\phi_{\mu_*\sigma_*}$ (broken lines) for the Gamma Distribution α = 1.0

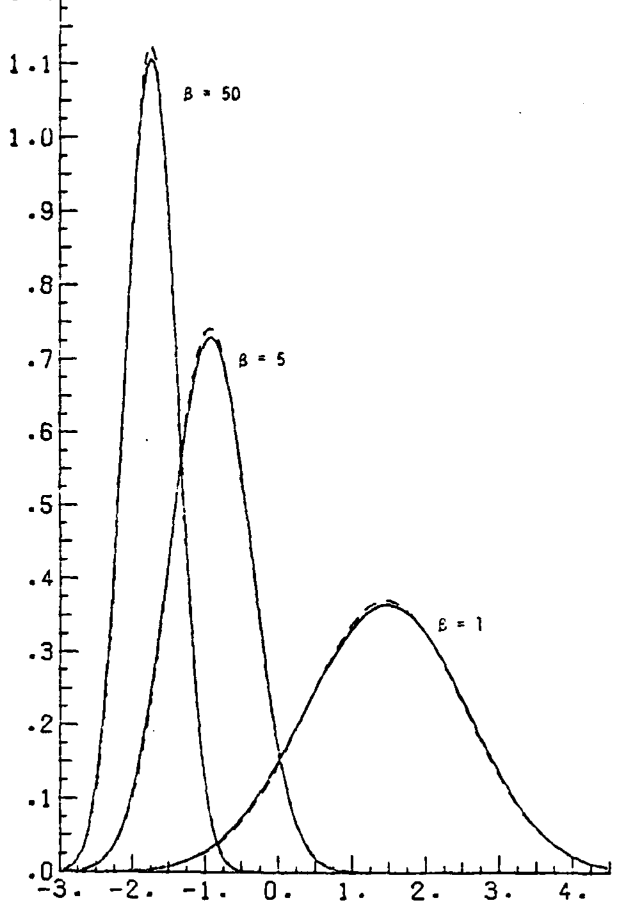
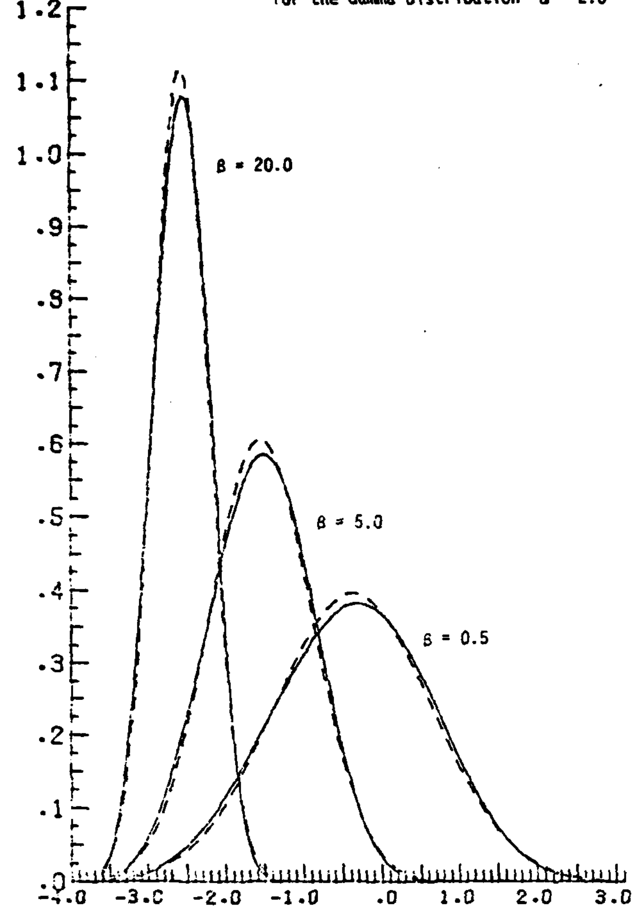
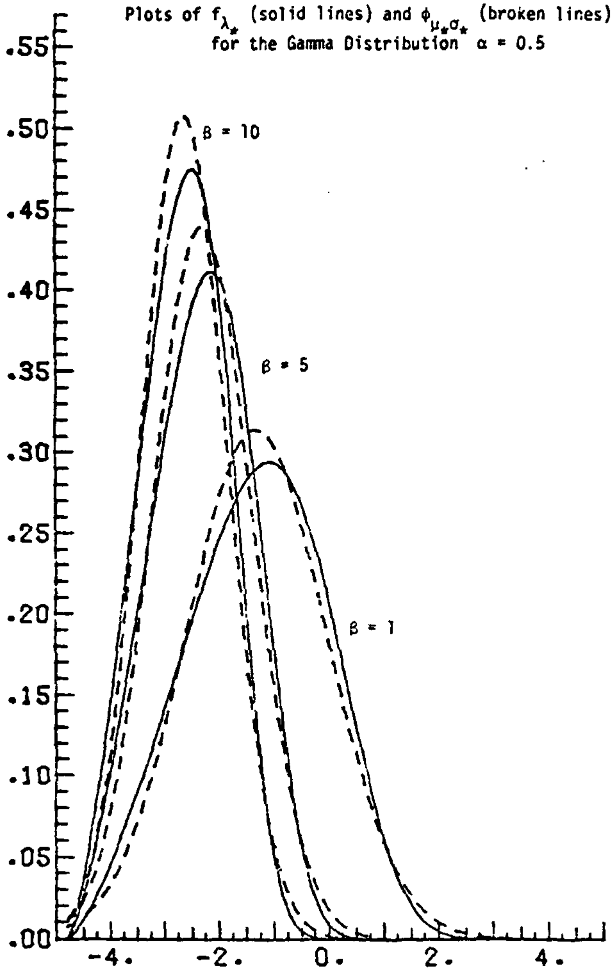


Figure 2



Plots of f_{λ_*} (solid lines) and $\phi_{\mu_*\sigma_*}$ (broken lines) for the Gamma Distribution $\alpha = 2.0$

Figure 3



Plots of f_{λ_*} (solid lines) and $\phi_{\mu_*\sigma_*}$ (broken lines) for the Gamma Distribution $\alpha = 0.5$

Figure 4

G(λ) for the Inverse Gaussian Distribution

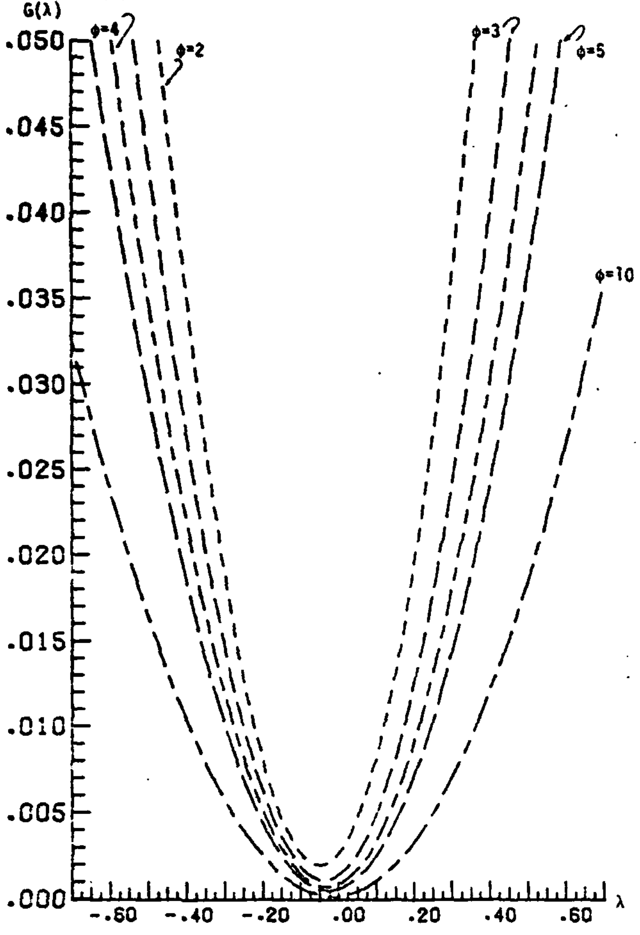


Figure 5

Plots of f_{λ_*} (solid line) and $\varphi_{\mu_*\sigma_*}$ (broken line) for the Inverse Gaussian Distribution $\alpha = 2, \ \mu = 1$

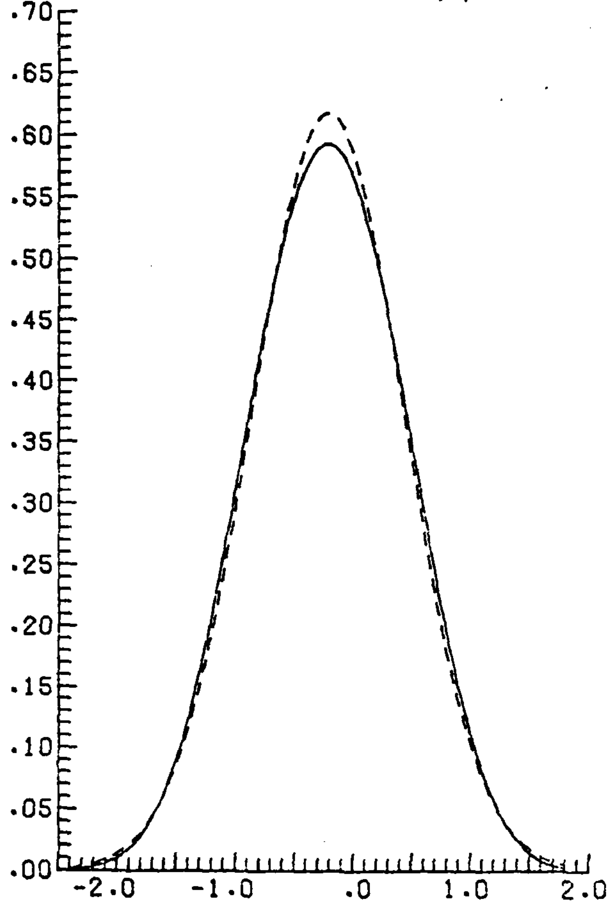


Figure 6

Plots of f_{λ_*}, $\phi_{\mu_*\sigma_*}$ and g, for the Pareto Distribution

$\alpha = 1$, $c = 1$

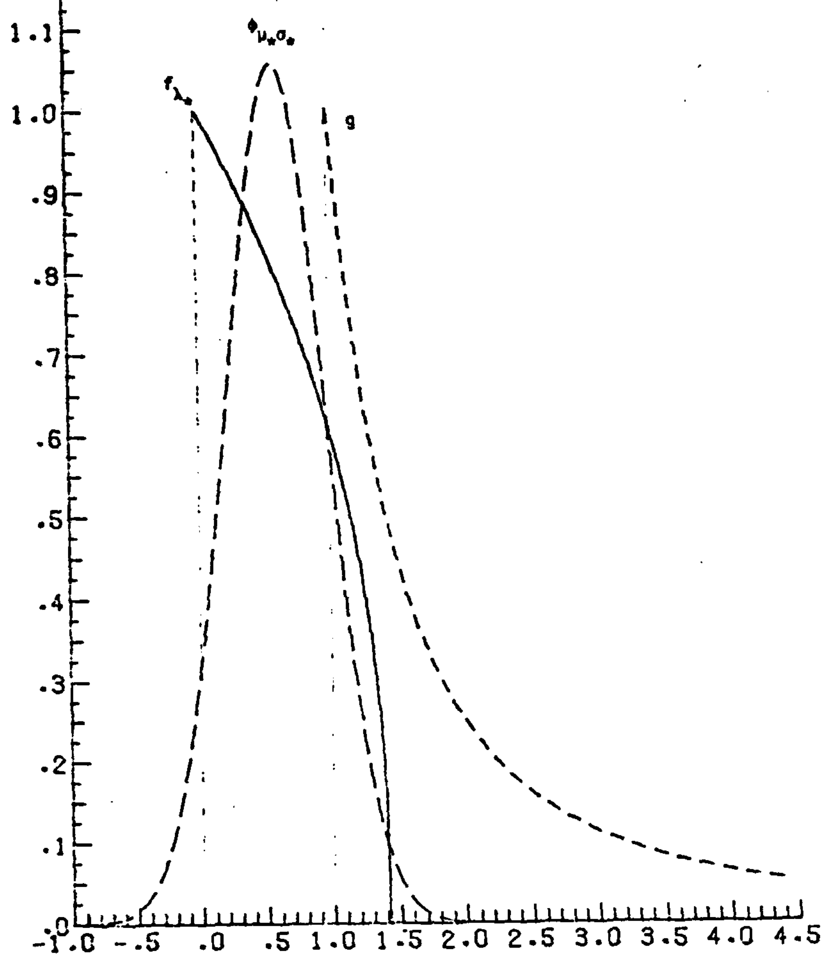


Figure 7

Level Curves for the Bivariate Gamma Distribution (broken curves) and the Normal Distribution (solid curves)

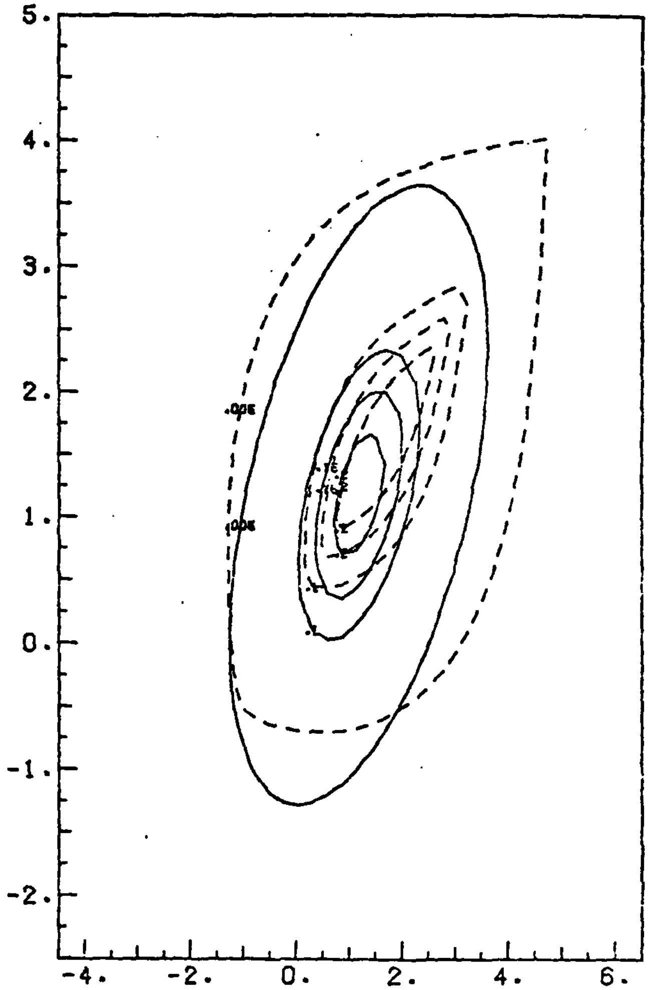


Figure 8

REFERENCES

Abramowitz, M. and Stegun, I.A. (1964). Handbook of Mathematical Functions, National Bureau of Standards, Applied Mathematics Series No. 55.

Berk, R.H. (1966). "Limiting Behavior of Posterior Distributions when the Model is Incorrect." Ann. Math. Statist., 37, 51-8.

Box, G.E.P. and Cox, D.R. (1964). "An Analysis of Transformations." J. Roy. Statist. Soc., Ser. B, 26, 211-43, discussion 244-52.

Draper, N.R. and Cox, D.R. (1969). "On Distributions and their Transformation to Normality." J. Roy. Statist. Soc., Ser. B, 31, 472-6.

Fuchs, C. (1978). "On Test Sizes in Linear Models for Transformed Variables." Technometrics, 20, 291-99.

Hernandez, F. (1978). The Large Sample Behavior of Transformations to Normal or Exponential Distributions. Ph.D. dissertation, University of Wisconsin at Madison.

Hinkley, D.V. (1975). "On Power Transformations to Symmetry." Biometrika, 62, 101-11.

Kullback, S. (1968). Information Theory and Statistics, New York: Dover Publications, Inc.

Rubin, H. (1956). "Uniform Convergence of Random Functions with Applications to Statistics." Ann. Math. Statist., 27, 200-203.

Tukey, J.W. (1977). Exploratory Data Analysis, Reading: Addison-Wesley Publishing Company.

Tweedie, M.C.K. (1957). "Statistical Properties of Inverse Gaussian Distributions, I." Ann. Math. Statist., 28, 362-377.

Watson, G.W. (1964). "A Note on Maximum Likelihood." Sankhyā, 26, A, 303-304.

Whitmore, G.A. and Yalovsky, M. (1978). "A Normalizing Logarithmic Transformation for Inverse Gaussian Random Variables." Technometrics, 20, 207-208.

Wilson, E.B. and Hilferty, M.M. (1931). "The Distribution of Chi-square." Proc. Nat. Acad. Sci. U.S.A., 17, 684-688.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report # 545	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) THE LARGE SAMPLE BEHAVIOR OF TRANSFORMATIONS TO NORMALITY		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Fabián Hernández Richard A. Johnson		8. CONTRACT OR GRANT NUMBER(s) ONR Grant No. N00014-78-C-0722
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics University of Wisconsin Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research 800 N. Quincy Street Arlington, VA 22217		12. REPORT DATE January 1979
		13. NUMBER OF PAGES 41 pages
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited.

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Transformations
Kullback Liebler Information

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We investigate the large sample behavior of both the classical and Bayesian procedures for selecting a transformation to normality. The study of the large sample behavior clearly reveals the role played by the assumptions leading to the Box and Cox procedures. Based on our large sample results, we introduce an information number approach for transforming a known distribution to near normality. This latter procedure provides bench marks for the maximum possible amount of improvement through power transformations. We illustrate our procedure with three examples. Finally, we generalize our procedure to random vectors and linear models situations

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)